

“© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

Taking A Closer Look at Domain Shift: Category-level Adversaries for Semantics Consistent Domain Adaptation

Yawei Luo^{1,2}, Liang Zheng⁵, Tao Guan^{1,6}, Junqing Yu^{1,4}*, Yi Yang^{2,3}

¹School of Computer Science & Technology, Huazhong University of Science & Technology

²CAI, University of Technology Sydney ³Baidu Research

⁴Center of Network and Computation, Huazhong University of Science & Technology

⁵Research School of Computer Science, Australian National University ⁶Farsee2 Tech. Co.

Abstract

We consider the problem of unsupervised domain adaptation in semantic segmentation. A key in this campaign consists in reducing the domain shift, i.e., enforcing the data distributions of the two domains to be similar. One of the common strategies is to align the marginal distribution in the feature space through adversarial learning. However, this global alignment strategy does not consider the category-level joint distribution. A possible consequence of such global movement is that some categories which are originally well aligned between the source and target may be incorrectly mapped, thus leading to worse segmentation results in target domain. To address this problem, we introduce a category-level adversarial network, aiming to enforce local semantic consistency during the trend of global alignment. Our idea is to take a close look at the category-level joint distribution and align each class with an adaptive adversarial loss. Specifically, we reduce the weight of the adversarial loss for category-level aligned features while increasing the adversarial force for those poorly aligned. In this process, we decide how well a feature is category-level aligned between source and target by a co-training approach. In two domain adaptation tasks, i.e., *GTA5* \rightarrow *Cityscapes* and *SYNTHIA* \rightarrow *Cityscapes*, we validate that the proposed method matches the state of the art in segmentation accuracy.

1. Introduction

Semantic segmentation aims to assign each pixel of a photograph to a semantic class label. Currently, the achievement is at the price of large amount of dense pixel-level

*Corresponding author (yjqing@hust.edu.cn).

This work was done when Yawei Luo (royalvane@hust.edu.cn) was a visiting student at University of Technology Sydney. Part of this work was done when Yi Yang (yee.i.yang@gmail.com) was visiting Baidu Research during his Professional Experience Program. The code is publicly available at <https://github.com/RoyalVane/CLAN>.

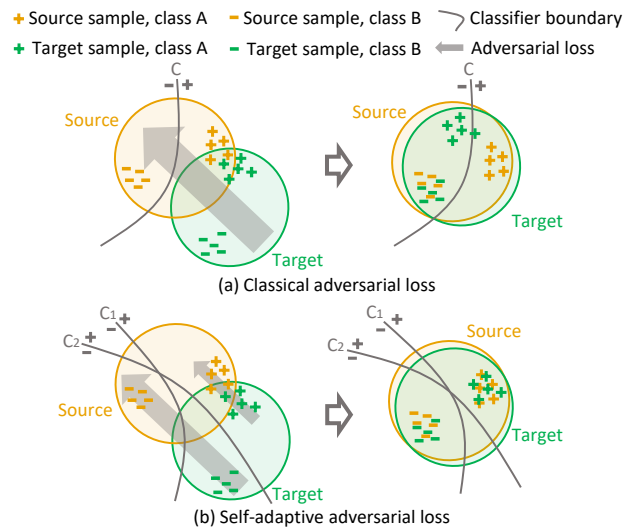


Figure 1. (Best viewed in color.) Illustration of traditional and the proposed adversarial learning. The size of the solid gray arrow represents the weight of the adversarial loss. (a) Traditional adversarial learning ignores the semantic consistency when pursuing the marginal distribution alignment. As a result, the global movement might cause the well-aligned features (class A) to be mapped onto different joint distributions (negative transfer). (b) The proposed self-adaptive adversarial learning reweights the adversarial loss for each feature by a local alignment score. Our method reduces the influence of the adversaries when discovers a high semantic alignment score on a feature, and vice versa. As is shown, the proposed strategy encourages a category-level joint distribution alignment for both class A and class B.

annotations obtained by expensive human labor [4, 23, 27]. An alternative would be resorting to simulated data, such as computer generated scenes [31, 32], so that unlimited amount of labels are made available. However, models trained with the simulated images do not generalize well to realistic domains. The reason lies in the different data distributions of the two domains, typically known as do-

main shift [37]. To address this issue, domain adaptation approaches [35, 41, 14, 46, 17, 16, 13, 48] are proposed to bridge the gap between the source and target domains. A majority of recent methods [26, 24, 40, 43, 42] aim to align the feature distributions of different domains. Works along this line are based on the theoretical insights in [1] that minimizing the divergence between domains lowers the upper bound of error on the target domain. Among this cohort of domain adaptation methods, a common and pivotal step is minimizing some distance metric between the source and target feature distributions [24, 40]. Another popular choice, which borrows the idea from adversarial learning [10], is to minimize the accuracy of domain prediction. Through a min-max game between two adversarial networks, the generator is trained to produce features that confuse the discriminator while the latter is required to correctly classify which domain the features are generated from.

Although the works along the path of adversarial learning have led to impressive results [39, 15, 22, 19, 43, 36], they suffer from a major limitation: when the generator network can perfectly fool the discriminator, it merely aligns the global marginal distribution of the features in the two domains (*i.e.*, $P(F_s) \approx P(F_t)$, where F_s and F_t denote the features of source and target domain in latent space) while ignores the local joint distribution shift, which is closely related to the semantic consistency of each category (*i.e.*, $P(F_s, Y_s) \neq P(F_t, Y_t)$, where Y_s and Y_t denote the categories of the features). As a result, the *de facto* use of the adversarial loss may cause those target domain features, which are already well aligned to their semantic counterpart in source domain, to be mapped to an incorrect semantic category (negative transfer). This side effect becomes more severe when utilize a larger weight on the adversarial loss.

To address the limitation of the global adversarial learning, we propose a category-level adversarial network (CLAN), prioritizing category-level alignment which will naturally lead to global distribution alignment. The cartoon comparison of traditional adversarial learning and the proposed one is shown in Fig. 1. The key idea of CLAN is two-fold. First, we identify those classes whose features are already well aligned between the source and target domains, and protect this category-level alignment from the side effect of adversarial learning. Second, we identify the classes whose features are distributed differently between the two domains and increase the weight of the adversarial loss during training. In this process, we utilize co-training [47], which enables high-confidence predictions with two diverse classifiers, to predict how well each feature is semantically aligned between the source and target domains. Specifically, if the two classifiers give consistent predictions, it indicates that the feature is predictive and achieves good semantic alignment. In such case, we reduce the influence of the adversarial loss in order to encourage the network to gen-

erate invariant features that can keep semantic consistency between domains. On the contrary, if the predictions disagree with each other, which indicates that the target feature is far from being correctly mapped, we increase the weight of the adversarial loss on that feature so as to accelerate the alignment. Note that 1) Our adversarial learning scheme acts directly on the output space. By regarding the output predictions as features, the proposed method jointly promotes the optimization for both classifier and extractor; 2) Our method does not guarantee rigorous joint distribution alignment between domains. Yet, compared with marginal distribution alignment, our method can *map the target features closer (or no negative transfer at worst)* to the source features of the same categories. The main contributions are summarized below.

- By proposing to adaptively weight the adversarial loss for different features, we emphasize the importance of category-level feature alignment in reducing domain shift.
- Our results are on par with the state-of-the-art UDA methods on two transfer learning tasks, *i.e.*, GTA5 [31] \rightarrow Cityscapes [8] and SYNTHIA [32] \rightarrow Cityscapes.

2. Related Works

This section will focus on adversarial learning and co-training techniques for unsupervised domain adaptation, which form the two main motivations of our method.

Adversarial learning. Ben-David *et al.* [1] had proven that the adaptation loss is bounded by three terms, *e.g.*, the expect loss on source domain, the domain divergence, and the shared error of the ideal joint hypothesis on the source and target domain. Because the first term corresponds to the well-studied supervised learning problems and the third term is considered sufficiently low to achieve an accurate adaptation, the majority of recent works lay emphasis on the second term. Adversarial adaptation methods are good examples of this type of approaches and can be investigated on different levels. Some methods focus on the distribution shift in the latent feature space [26, 39, 15, 22, 19, 43, 36]. In an example, Hoffman *et al.* [15] appended category statistic constraints to the adversarial model, aiming to improve semantic consistency in target domain. Other methods address the adaption problem on the pixel level [21, 3], which relate to the style transfer approaches [49, 7] to make images indistinguishable across domains. A joint consideration of pixel and feature level domain adaptation is studied in [14]. Besides alignment in the bottom feature layers, Tsai *et al.* [41] found that aligning directly the output space is more effective in semantic segmentation. Domain adaptation in the output space enables the joint optimization for both prediction and representation, so our method utilizes this advantage.

Co-training. Co-training [47] belongs to multi-view learning in which learners are trained alternately on two distinct views with confident labels from the unlabeled data. In UDA, this line of methods [44, 5, 33, 25] are able to assign pseudo labels to unlabeled samples in the target domain, which enables direct measurement and minimization the classification loss on target domain. In general, co-training enforces the two classifiers to be diverse in the learned parameters, which can be achieved via dropout [34], consensus regularization [35] or parameter diverse [44], *etc.* Similar to co-training, tri-training keeps the two classifiers producing pseudo labels and uses these pseudo labels to train an extra classifier [33, 44]. Apart from assigning pseudo labels to unlabeled data, Saiko *et al.* [34, 35] maximized the consensus of two classifiers for domain adaptation.

Our work does not follow the strategy of global feature alignment [41, 15, 39] or classifiers consensus maximization [34, 35]. Instead, category-level feature alignment is enforced through co-training. To our knowledge, we are making an early attempt to adaptively weight the adversarial loss for features in segmentation task according to the local alignment situation.

3. Method

3.1. Problem Settings

We focus on the problem of unsupervised domain adaptation (UDA) in semantic segmentation, where we have access to the source data X_S with pixel-level labels Y_S , and the target data X_T without labels. The goal is to learn a model G that can correctly predict the pixel-level labels for the target data X_T . Traditional adversaries-based networks (TAN) consider two aspects for domain adaptation. First, these methods train a model G that distills knowledge from labeled data in order to minimize the segmentation loss in the source domain, formalized as a fully supervised problem:

$$\mathcal{L}_{seg}(G) = E[\ell(G(X_S), Y_S)], \quad (1)$$

where $E[\cdot]$ denotes statistical expectation and $\ell(\cdot, \cdot)$ is an appropriate loss function, such as multi-class cross entropy.

Second, adversaries-based UDA methods also train G to learn domain-invariant features by confusing a domain discriminator D which is able to distinguish between samples of the source and target domains. This property is achieved by minimaxing an adversarial loss:

$$\begin{aligned} \mathcal{L}_{adv}(G, D) = & -E[\log(D(G(X_S)))] \\ & -E[\log(1 - D(G(X_T)))] . \end{aligned} \quad (2)$$

However, as mentioned above, there is a major limitation for traditional adversarial learning methods: even under perfect alignment in marginal distribution, there might be

the negative transfer that causes the samples from different domains but of the same class label to be mapped farther away in the feature space. In some cases, some classes are already aligned between domains, but the adversarial loss might deconstruct the existing local alignment when pursuing the global marginal distribution alignment. In this paper, we call this phenomenon “lack of semantic consistency”, which is a critical cause of performance degradation.

3.2. Network Architecture

Our network architecture is illustrated in Fig. 2. It is composed of a generator G and a discriminator D . G can be any FCN-based segmentation network [38, 23, 4] and D is a CNN-based binary classifier with a fully-convolutional output [10]. As suggested in the standard co-training algorithm [47], generator G is divided into feature extractor E and two classifiers C_1 and C_2 . E extracts features from input images; C_1 and C_2 classify features generated from E into one of the pre-defined semantic classes, such as car, tree and road. Following the co-training practice, we enforce the weights of C_1 and C_2 to be diverse through a cosine distance loss. This will provide us with the distinct views / classifiers to make semantic predictions for each feature. The final prediction map p is obtained by summing up the two diverse prediction tensors $p^{(1)}$ and $p^{(2)}$ and we call p an *ensemble prediction*.

Given a source domain image $x_s \in X_S$, feature extractor E outputs a feature map, which is input to classifiers C_1 and C_2 to yield the pixel-level ensemble prediction p . On the one hand, p is used to calculate a segmentation loss under the supervision of the ground-truth label $y_s \in Y_S$. On the other hand, p is input to D to generate an adversarial loss.

Given a target domain image $x_t \in X_T$, we also forward it to G and obtain an ensemble prediction p . Different from the source data flow, we additionally generate a discrepancy map out of $p^{(1)}$ and $p^{(2)}$, denoted as $\mathcal{M}(p^{(1)}, p^{(2)})$, where $\mathcal{M}(\cdot, \cdot)$ denotes some proper distance metric to measure the element-wise discrepancy between $p^{(1)}$ and $p^{(2)}$. When using the cosine distance as an example, $\mathcal{M}(p^{(1)}, p^{(2)})$ forms a $1 \times H \times W$ shaped tensor with the $(i_{th} \in H, j_{th} \in W)$ element equaling to $(1 - \cos(p_{i,j}^{(1)}, p_{i,j}^{(2)}))$. Once D produces an adversarial loss map \mathcal{L}_{adv} , an element-wise multiplication is performed between \mathcal{L}_{adv} and $\mathcal{M}(p^{(1)}, p^{(2)})$. As a result, the final adaptive adversarial loss on a target sample takes the form as $\sum_{i=1}^H \sum_{j=1}^W (1 - \cos(p_{i,j}^{(1)}, p_{i,j}^{(2)})) \times \mathcal{L}_{adv_{i,j}}$, where $\{i, j\}$ traverses over all the pixels on the map. In this manner, each pixel on the segmentation map is differently weighted *w.r.t* the adversarial loss.

3.3. Training Objective

The proposed network is featured by three loss functions, *i.e.*, the *segmentation loss*, the *weight discrepancy loss* and the *self-adaptive adversarial loss*. Given an image $x \in X_S$

→ Source flow → Target flow ↔ Weight discrepancy Σ Tensor sum M Distance metric \otimes Element-wise product

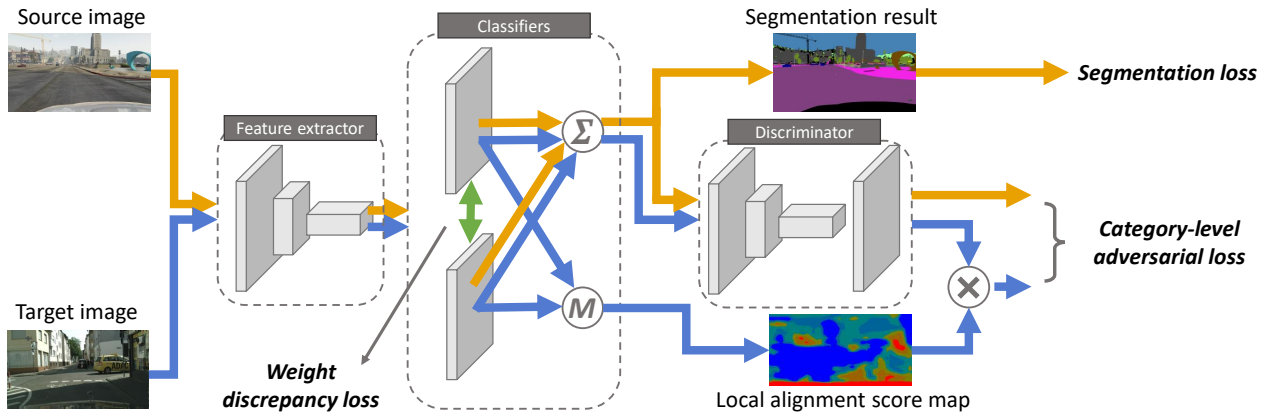


Figure 2. Overview of the proposed category-level adversarial network. It consists of a feature extractor E , two classifiers C_1 and C_2 , and a discriminator D . C_1 and C_2 are fed with the deep feature map extracted from E and predict semantic labels for each pixel from diverse views. In source flow, the sum of the two prediction maps is used to calculate a segmentation loss as well as an adversarial loss from D . In target flow, the sum of the two prediction maps is forwarded to D to produce a *raw* adversarial loss map. Additionally, we adopt the discrepancy of the two prediction maps to produce a local alignment score map. This map evaluates the category-level alignment degree of each feature and is used to adaptively weight the raw adversarial loss map.

of shape $3 \times H \times W$ and a label map $y \in Y_S$ of shape $C \times H \times W$ where C is the number of semantic classes, the segmentation loss (multi-class cross-entropy loss) can be concretized from Eq. 1 as

$$\mathcal{L}_{seg}(G) = \sum_{i=1}^{H \times W} \sum_{c=1}^C -y_{ic} \log p_{ic}, \quad (3)$$

where p_{ic} denotes the predicted probability of class c on pixel i . y_{ic} denotes the ground truth probability of class c on the pixel i . If pixel i belongs to class c , $y_{ic} = 1$, otherwise $y_{ic} = 0$.

For the second loss, as suggested in the standard co-training algorithm [47], the two classifiers C_1 and C_2 should have possibly diverse parameters in order to provide two different views on a feature. Otherwise, the training degenerates to self-training. Specifically, we enforce divergence of the weights of the convolutional layers of the two classifiers by *minimizing* their cosine similarity. Therefore, we have the following weight discrepancy loss:

$$\mathcal{L}_{weight}(G) = \frac{\vec{w}_1 \cdot \vec{w}_2}{\|\vec{w}_1\| \|\vec{w}_2\|}, \quad (4)$$

where \vec{w}_1 and \vec{w}_2 are obtained by flattening and concatenating the weights of the convolution filters of C_1 and C_2 .

Third, we adopt the discrepancy between the two predictions $p^{(1)}$ and $p^{(2)}$ as an indicator to weight the adversarial loss. The self-adaptive adversarial loss can be extended from

the traditional adversarial loss (Eq. 2) as

$$\mathcal{L}_{adv}(G, D) = -E[\log(D(G(X_S)))] - E[(\lambda_{local} \mathcal{M}(p^{(1)}, p^{(2)}) + \epsilon) \log(1 - D(G(X_T)))] , \quad (5)$$

where $p^{(1)}$ and $p^{(2)}$ are predictions made by C_1 and C_2 , respectively, $\mathcal{M}(\cdot, \cdot)$ denotes the cosine distance, and λ_{local} controls the adaptive weight for adversarial loss. Note that in Eq. 5, to stabilize the training process, we add a small number ϵ to the self-adaptive weight.

With the above loss terms, the overall loss function of our approach can be written as

$$\mathcal{L}_{CLAN}(G, D) = \mathcal{L}_{seg}(G) + \lambda_{weight} \mathcal{L}_{weight}(G) + \lambda_{adv} \mathcal{L}_{adv}(G, D), \quad (6)$$

where λ_{weight} and λ_{adv} denote the hyper parameters that control the relative importance of the three losses. The training objective of CLAN is

$$G^*, D^* = \arg \min_G \max_D \mathcal{L}_{CLAN}(G, D). \quad (7)$$

We solve Eq. 7 by alternating between optimizing G and D until $\mathcal{L}_{CLAN}(G, D)$ converges.

3.4. Analysis

The major difference between the proposed framework and traditional adversarial learning consists in two aspects: the discrepancy loss and the category-level adversarial loss. Accordingly, analysis will focus on the two differences.

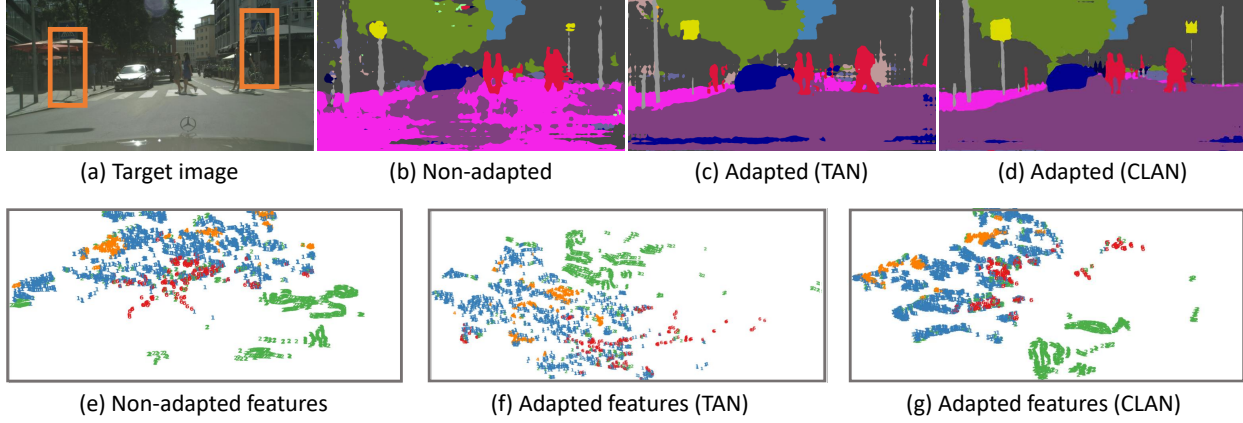


Figure 3. A contrastive analysis of CLAN and traditional adversarial network (TAN). (a): A target image, and we focus on the poles and traffic signs in orange boxes. (b): A non-adapted segmentation result. Although the global segmentation result is poor, the poles and traffic signs can be correctly segmented. It indicates that some classes are originally aligned between domains, even without any domain adaptation. (c): Adapted result of TAN, in which a decent segmentation map is produced but poles and traffic signs are poorly segmented. The reason is that the global alignment strategy tends to assign a *conservative prediction* to a feature and would lead some features to be predicted to other *prevalent classes* [11, 18], thus causing those infrequent features being negatively transferred. (d): Adapted result from CLAN. CLAN reduces the weight of adversarial loss for those aligned features. As a result, the original well-segmented class are well preserved. We then map the high-dimensional features of (b), (c) and (d) to a 2-D space with t-SNE [29] shown in (e), (f) and (g). The comparison of feature distributions further proves that CLAN can enforce category-level alignment during the trend of global alignment. (For a clear illustration, we only show 4 related classes, *i.e.*, building in blue, traffic sign in orange, pole in red and vegetation in green.)

First, the discrepancy (co-training) loss encourages E to learn domain-invariant semantics instead of the domain specific elements such as illumination. In our network, classifiers C_1 and C_2 1) are encouraged to capture possibly different characteristics of a feature, which is ensured by the discrepancy loss, and 2) are enforced to make the same prediction of any E output (no matter the source or target), which is required by the segmentation loss and the adversarial loss. The two forces actually require that E should capture the essential aspect of a pixel across the source and target domains, which, as we are aware of, is the pure semantics of a pixel, *i.e.*, the domain-invariant aspect of a pixel. Without the discrepancy loss (co-training), force 1) is missing, and there is a weaker requirement for E to learn domain-invariant information. On the other side, in our *simulated* \rightarrow *real* task, the two domains vary a lot at *visual level*, but overlap at *semantic level*. If C_1 and C_2 are input with *visual-level features* from E , their predictions should be inaccurate in target domain and tend to be different, which will be *punished by large adversarial losses*. As a result, once our algorithm converges, C_1 and C_2 will be input with *semantic-level features* instead of *visual-level features*. That is, E is encouraged to learn domain-invariant semantics. Therefore, the discrepancy loss serves as an implicit contributing factor for the improved adaptation ability.

Second, in our major contribution, we extend the traditional adversarial loss with an adaptive weight $[\lambda_{local}\mathcal{M}(p^{(1)}, p^{(2)}) + \epsilon]$. On the one hand, when $\mathcal{M}(p^{(1)}, p^{(2)})$ is large, feature maps of the same class do

not have similar joint distributions between two domains: they suffer from the semantic inconsistency. Therefore, the weights are such assigned as to encourage G to fool D mainly on features that suffer from domain shift. On the other hand, when $\mathcal{M}(p^{(1)}, p^{(2)})$ is small, the joint distribution would have a large overlap across domains, indicating that the semantic inconsistency problem is not severe. Under this circumstance, G tends to ignore the adversarial punishment from D . From the view of D , the introduction of the adaptive weight encourages D to distill more knowledge from examples suffering from semantic inconsistency rather than those well-aligned classes. As a result, CLAN is able to improve category-level alignment degree in adversarial training. This could be regarded as an explicit contributing factor for the adaptation ability. We additionally give a contrastive analysis between traditional adversarial network (TAN) and CLAN on their adaptation result in Fig. 3.

4. Experiment

4.1. Datasets

We evaluate CLAN together with several state-of-the-art algorithms on two adaptation tasks, *e.g.*, SYNTHIA [32] \rightarrow Cityscapes [8] and GTA5 [31] \rightarrow Cityscapes. Cityscapes is a real-world dataset with 5,000 street scenes. We use Cityscapes as the target domain. GTA5 contains 24,966 high-resolution images compatible with the Cityscapes annotated classes. SYNTHIA contains 9400 synthetic images. We use SYNTHIA or GTA5 as the source domain.

Table 1. Adaptation from GTA5 [31] to Cityscapes [8]. We present per-class IoU and mean IoU. “V” and “R” represent the VGG16-FCN8s and ResNet101 backbones, respectively. “ST” and “AT” represent two lines of method, *i.e.*, self training- and adversarial learning-based DA. We highlight the best result in each column in **bold**. To clearly showcase the effect of CLAN on infrequent classes, we highlight these classes in **blue**. *Gain* indicates the mIoU improvement over using the source only.

GTA5 → Cityscapes																							
	Arch.	Meth.	road	side.	buil.	wall	fence	pole	light	sign	vege.	terr.	sky	pers.	rider	car	truck	bus	train	motor	bike	mIoU	gain
Source only	V	-	64.0	22.1	68.6	13.3	8.7	19.9	15.5	5.9	74.9	13.4	37.0	37.7	10.3	48.2	6.1	1.2	1.8	10.8	2.9	24.3	—
CBST [50]	V	ST	90.4	50.8	72.0	18.3	9.5	27.2	28.6	14.1	82.4	25.1	70.8	42.6	14.5	76.9	5.9	12.5	1.2	14.0	28.6	36.1	11.8
Source only	V	-	25.9	10.9	50.5	3.3	12.2	25.4	28.6	13.0	78.3	7.3	63.9	52.1	7.9	66.3	5.2	7.8	0.9	13.7	0.7	24.9	—
MCD [35]	V	AT	86.4	8.5	76.1	18.6	9.7	14.9	7.8	0.6	82.8	32.7	71.4	25.2	1.1	76.3	16.1	17.1	1.4	0.2	0.0	28.8	3.9
Source only	V	-	18.1	6.8	64.1	7.3	8.7	21.0	14.9	16.8	45.9	2.4	64.4	41.6	17.5	55.3	8.4	5.0	6.9	4.3	13.8	22.3	—
CDA [45]	V	AT	74.9	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9	6.6
Source only	V	-	26.0	14.9	65.1	5.5	12.9	8.9	6.0	2.5	70.0	2.9	47.0	24.5	0.0	40.0	12.1	1.5	0.0	0.0	0.0	17.9	—
FCNs in the wild [15]	V	AT	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1	9.2
CyCADA (feature) [14]	V	AT	85.6	30.7	74.7	14.4	13.0	17.6	13.7	5.8	74.6	15.8	69.9	38.2	3.5	72.3	16.0	5.0	0.1	3.6	0.0	29.2	11.3
Baseline (TAN) [41]	V	AT	87.3	29.8	78.6	21.1	18.2	22.5	21.5	11.0	79.7	29.6	71.3	46.8	6.5	80.1	23.0	26.9	0.0	10.6	0.3	35.0	17.1
CLAN	V	AT	88.0	30.6	79.2	23.4	20.5	26.1	23.0	14.8	81.6	34.5	72.0	45.8	7.9	80.5	26.6	29.9	0.0	10.7	0.0	36.6	18.7
Source only	R	-	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6	—
Baseline (TAN) [41]	R	AT	86.5	25.9	79.8	22.1	20.0	23.6	33.1	21.8	81.8	25.9	75.9	57.3	26.2	76.3	29.8	32.1	7.2	29.5	32.5	41.4	4.8
CLAN	R	AT	87.0	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28.0	76.2	33.1	36.7	6.7	31.9	31.4	43.2	6.6

Table 2. Adaptation from SYNTHIA [32] to Cityscapes [8]. We present per-class IoU and mean IoU for evaluation. CLAN and state-of-the-art domain adaptation methods are compared. For each backbone, the best accuracy is highlighted in **bold**. To clearly showcase the effect of CLAN on infrequent classes, we highlight these classes in **blue**. *Gain* indicates the mIoU improvement over using the source only.

SYNTHIA → Cityscapes																	
	Arch.	Meth.	road	side.	bul.	light	sign	vege.	sky	pers.	rider	car	bus	motor	bike	mIoU	gain
Source only	V	-	17.2	19.7	47.3	3.0	9.1	71.8	78.3	37.6	4.7	42.2	9.0	0.1	0.9	26.2	—
CBST [50]	V	ST	69.6	28.7	69.5	11.9	13.6	82.0	81.9	49.1	14.5	66.0	6.6	3.7	32.4	36.1	9.9
Source only	V	-	6.4	17.7	29.7	0.0	7.2	30.3	66.8	51.1	1.5	47.3	3.9	0.1	0.0	20.2	—
FCNs in the wild [15]	V	AT	11.5	19.6	30.8	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	22.9	2.7
Cross-city [6]	V	AT	62.7	25.6	78.3	1.2	5.4	81.3	81.0	37.4	6.4	63.5	16.1	1.2	4.6	35.7	15.2
Baseline (TAN) [41]	V	AT	78.9	29.2	75.5	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	37.6	17.4
CLAN	V	AT	80.4	30.7	74.7	1.4	8.0	77.1	79.0	46.5	8.9	73.8	18.2	2.2	9.9	39.3	19.1
Source only	R	-	55.6	23.8	74.6	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	38.6	—
Baseline (TAN) [41]	R	AT	79.2	37.2	78.8	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	45.9	7.3
CLAN	R	AT	81.3	37.0	80.1	16.1	13.7	78.2	81.5	53.4	21.2	73.0	32.9	22.6	30.7	47.8	9.2

4.2. Implementation Details

We use PyTorch for implementation. We utilize the DeepLab-v2 [4] framework with ResNet-101 [12] pre-trained on ImageNet [9] as our source-only backbone for network G . We use the single layer adversarial DA method proposed in [41] as the TAN baseline. For co-training, we duplicate two copies of the last classification module and arrange them in parallel after the feature extractor, as illustrated in Fig. 2. For a fair comparison to those methods with the VGG backbone, we also apply CLAN on VGG-16 based FCN8s [23]. For network D , we adopt a similar structure with [30], which consists of 5 convolution layers with kernel 4×4 with channel numbers $\{64, 128, 256, 512, 1\}$ and stride of 2. Each convolution layer is followed by a

Leaky-ReLU [28] parameterized by 0.2 except the last layer. Finally, we add an up-sampling layer to the last layer to rescale the output to the size of the input map, in order to match the size of local alignment score map. During training, we use SGD [2] as the optimizer for G with a momentum of 0.9, while using Adam [20] to optimize D with $\beta_1 = 0.9$, $\beta_2 = 0.99$. We set both optimizers a weight decay of $5e - 4$. For SGD, the initial learning rate is set to $2.5e - 4$ and decayed by a poly learning rate policy, where the initial learning rate is multiplied by $(1 - \frac{iter}{max.iter})^{power}$ with $power = 0.9$. For Adam, we initialize the learning rate to $5e - 5$ and fix it during the training. We train the network for a total of 100k iterations. We use a crop of 512×1024 during training, and for evaluation we up-sample the prediction map by a factor of 2 and then evaluate mIoU. In our best

model, the hyper-parameters λ_{weight} , λ_{adv} , λ_{local} and ϵ are set to 0.01, 0.001, 40 and 0.4 respectively.

4.3. Comparative Studies

We present the adaptation results on task GTA5 \rightarrow Cityscapes in Table 1 with comparisons to the state-of-the-art domain adaptation methods [35, 45, 15, 14, 41, 50]. We observe that CLAN significantly outperforms the source-only segmentation method by +18.7% on VGG-16 and +6.6% on ResNet-101. Besides, CLAN also outperforms the state-of-the-art methods, which improves the mIOU by over +7% compared with MCD [35], CDA [45] and CyCADA [14]. Compared to traditional adversarial network (TAN) in the output space [41], CLAN brings over +1.6% improvement in mIOU in both architectures of VGG-16 and ResNet-101. In some infrequent classes which are prone to suffer from the side effect of global alignment, *e.g.*, fence, traffic light and pole, CLAN can significantly outperform TAN. Besides, we also compare CLAN with the self training-based methods, among which CBST [50] is the current state-of-the-art one. This series of explicit methods usually achieve higher mIOU than the implicit feature alignment. While in our experiment, we find that CLAN is on par with CBST. Some qualitative segmentation examples can be viewed in Fig. 5.

Table 2 provides the comparative results on the task SYNTHIA \rightarrow Cityscapes. On VGG-16, our final model yields 39.3% in terms of mIOU, which significantly improves the non-adaptive segmentation result by 19.1%. Besides, CLAN outperforms the current state-of-art method [15] by 16.4% and [6] by 3.6%. On ResNet-101, CLAN brings 9.2% improvement to source only segmentation model. Compare to TAN [41], the use of adaptive adversarial loss also brings 1.9% gain in terms of mIOU. Likewise, CLAN is more effective on those infrequent classes which are prone to be negatively transferred, such as traffic light and sign, bringing over 3.2% improvement respectively. While on some prevalent classes, CLAN can also be par on with the baseline method. Note that on the “train” class, the improvement is not stable. This is due to the training samples that contain the “train” are very few. Finally, comparing with the self training-based method, CLAN outperforms CBST by 3.2% in terms of mIOU. These observations are in consistent with our t-SNE analysis in Fig. 3, which further verifies that CLAN can actually boost the category-level alignment in segmentation-based DA task.

4.4. Feature Distribution

To further verify that CLAN is able to decrease the negative transfer effect for those well-aligned features, we designed an experiment to take a closer look at the category-level alignment degree of each class. Specifically, we randomly select 1K source and 1K target images and calculate the cluster center distance (CCD) $\{d_1^e \dots d_n^e\}$ of features of

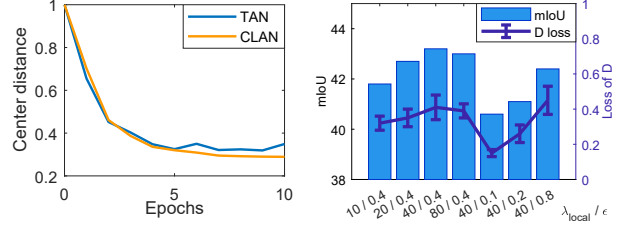


Figure 4. **Left:** Cluster center distance variation as training goes on. **Right:** Mean IoU (see bars & left y axis) and convergence performance (see lines & right y axis) variation when training with different λ_{local} and ϵ .

the same class between two domains, where $n = \#class$ and e is training epoch. d_i^e is normalized by d_i^e / d_i^0 (In this way, the CCD from the pre-trained model without any fine-tuning would be always normalized to 1). We report d_i^e in Fig. 4 (left subfigure, taking the class “wall” as an example). First, we observe as training goes on, d_i^e is monotonically decreasing in CLAN while not being monotone in TAN, suggesting CLAN prevents the well-aligned features from being incorrectly mapped. Second, d_i^e converges to a smaller value in CLAN than TAN, suggesting CLAN achieves better feature alignment at semantic level.

We further report the final CCD of each class in Fig. 6. We can observe that CLAN can achieve a smaller CCD in most cases, especially in those infrequent classes which are prone to be negatively transferred. These quantitative results, together with the qualitative t-SNE [29] analysis in Fig. 3, indicate that CLAN can preferably align the two domains in semantic level. Such category-aligned feature distribution usually makes the subsequent classification easier.

4.5. Parameter Studies

In this experiment, we aim to study two problems: 1) whether the adaptive adversarial loss would cause instability (vanishing gradient) during adversarial training and 2) how much the adaptive adversarial loss would effect the performance. For the problem 1), we utilize the loss of D to indicate the convergence performance and a stable adversarial training is achieved if D loss converges around 0.5. First, we test our model using $\lambda_{local} = 40$, with varying ϵ over a range $\{0.1, 0.2, 0.4, 0.8\}$. We do not use any ϵ larger than 0.8 since CLAN would degrade into TAN in that case. In the experiment, our model suffers from poor convergence when utilize a very small ϵ , *e.g.*, 0.1 or 0.2. It indicates that a proper choice of ϵ is between 0.2 and 0.8. Motivated by this observation, we then test our model using $\epsilon = 0.4$ with varying λ_{local} over a range $\{10, 20, 40, 80\}$. We observe that the convergence performance is not very sensitive to λ_{local} since the loss of D converges to proper values in all the cases. The best performance is achieved when using $\lambda_{local} = 40$ and $\epsilon = 0.4$. Besides, we observe that the adaptation performance of CLAN can steadily outperform TAN when using

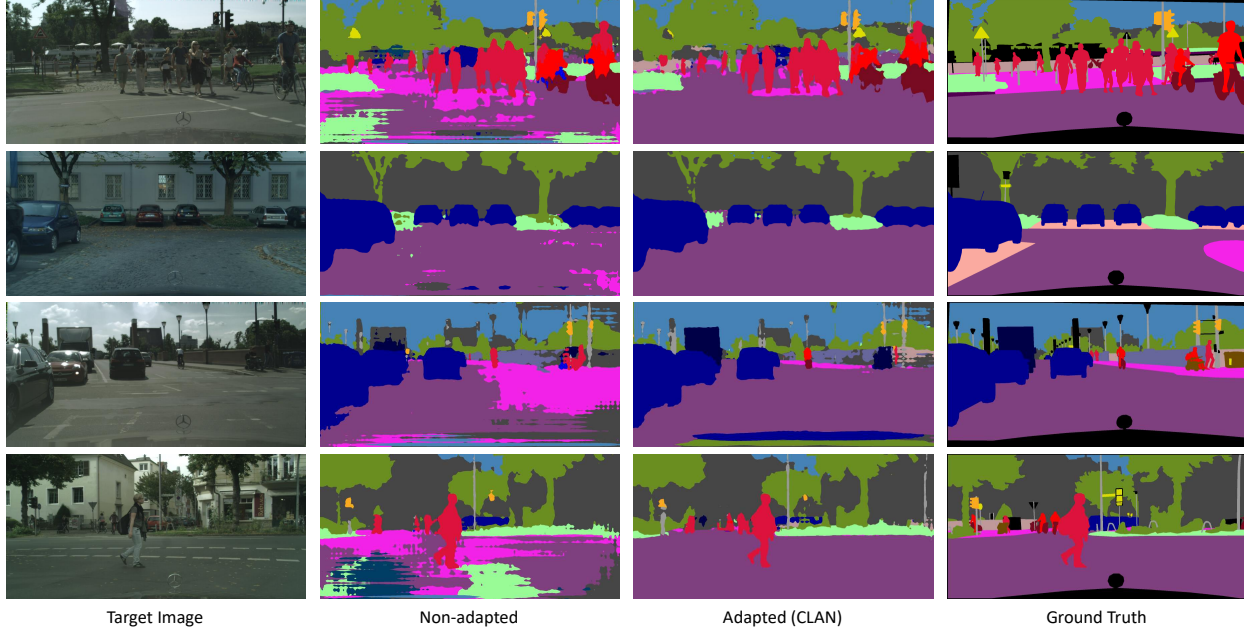


Figure 5. Qualitative results of UDA segmentation for GTA5 → Cityscapes. For each target image, we show the non-adapted (source only) result, adapted result with CLAN and the ground truth label map, respectively.

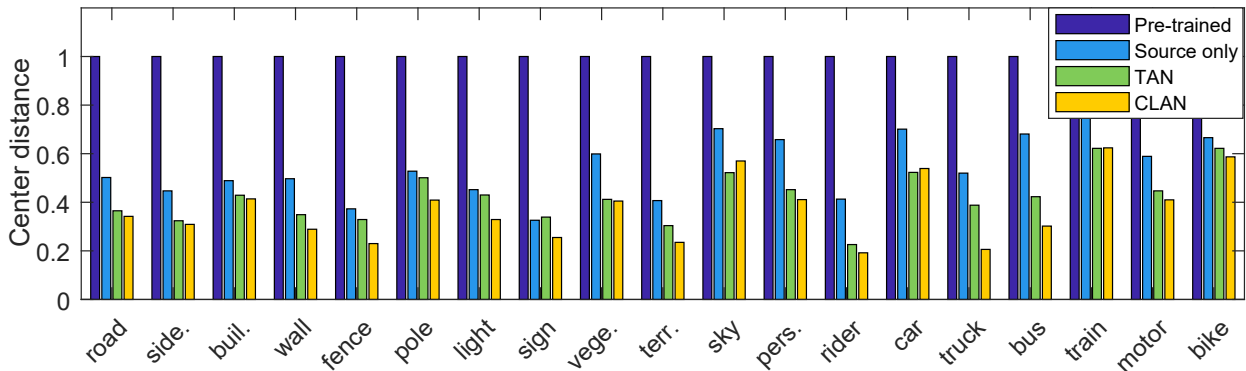


Figure 6. Quantitative analysis of the feature joint distributions. For each class, we show the distance of the feature cluster centers between source domain and target domain. These results are from 1) the model pre-trained on ImageNet [9] without any fine-tuning, 2) the model fine-tuned with source images only, 3) the adapted model using TAN and 4) the adapted model using CLAN, respectively.

parameters near the best value. We present the detailed performance variation in Fig. 4 (right subfigure). By comparing both the convergence and segmentation results under these different parameter settings, we can conclude that our proposed adaptive adversarial weight can significantly effect and improve the adaptation performance.

5. Conclusion

In this paper, we introduce the category-level adversarial network (CLAN), aiming to address the problem of semantic inconsistency incurred by global feature alignment during unsupervised domain adaptation (UDA). By taking a close look at the category-level data distribution, CLAN

adaptively weight the adversarial loss for each feature according to how well their category-level alignment is. In this spirit, each class is aligned with an adaptive adversarial loss. Our method effectively prevents the well-aligned features from being incorrectly mapped by the side effect of pure global distribution alignment. Experimental results validate the effectiveness of CLAN, which yields very competitive segmentation accuracy compared with state-of-the-art UDA approaches.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (No. 61572211).

References

- [1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.
- [2] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT’2010*, pages 177–186. Springer, 2010.
- [3] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 7, 2017.
- [4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.
- [5] M. Chen, K. Q. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *Advances in neural information processing systems*, pages 2456–2464, 2011.
- [6] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2011–2020. IEEE, 2017.
- [7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.
- [8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [13] Y. He, P. Liu, Z. Wang, Z. Hu, and Y. Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [14] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell. Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.
- [15] J. Hoffman, D. Wang, F. Yu, and T. Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.
- [16] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [17] G. Kang, L. Zheng, Y. Yan, and Y. Yang. Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–416, 2018.
- [18] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.
- [19] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.
- [20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [21] P. Li, X. Liang, D. Jia, and E. P. Xing. Semantic-aware grad-gan for virtual-to-real urban scene adaption. *arXiv preprint arXiv:1801.01726*, 2018.
- [22] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016.
- [23] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [24] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.
- [25] P. Luo, F. Zhuang, H. Xiong, Y. Xiong, and Q. He. Transfer learning from multiple source domains via consensus regularization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 103–112. ACM, 2008.
- [26] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang. Significance-aware information bottleneck for domain adaptive semantic segmentation. *arXiv preprint arXiv:1904.00876*, 2019.
- [27] Y. Luo, Z. Zheng, L. Zheng, T. Guan, J. Yu, and Y. Yang. Macro-micro adversarial network for human parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.
- [28] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.
- [29] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [30] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

- [31] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016.
- [32] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016.
- [33] K. Saito, Y. Ushiku, and T. Harada. Asymmetric tri-training for unsupervised domain adaptation. *arXiv preprint arXiv:1702.08400*, 2017.
- [34] K. Saito, Y. Ushiku, T. Harada, and K. Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017.
- [35] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. *arXiv preprint arXiv:1712.02560*, 2017.
- [36] S. Sankaranarayanan, Y. Balaji, A. Jain, S. N. Lim, and R. Chellappa. Unsupervised domain adaptation for semantic segmentation with gans. *arXiv preprint arXiv:1711.06969*, 2017.
- [37] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.
- [38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [39] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.
- [40] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, pages 443–450. Springer, 2016.
- [41] Y.-H. Tsai, W.-C. Hung, S. Schuster, K. Sohn, M.-H. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. *arXiv preprint arXiv:1802.10349*, 2018.
- [42] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4068–4076. IEEE, 2015.
- [43] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial discriminative domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2017.
- [44] J. Zhang, C. Liang, and C.-C. J. Kuo. A fully convolutional tri-branch network (fctn) for domain adaptation. *arXiv preprint arXiv:1711.03694*, 2017.
- [45] Y. Zhang, P. David, and B. Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 6, 2017.
- [46] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019.
- [47] Z.-H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005.
- [48] F. Zhu, L. Zhu, and Y. Yang. Sim-real joint reinforcement transfer for 3d indoor navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.
- [50] Y. Zou, Z. Yu, B. Vijaya Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018.